Developing LMF-XML Bilingual Dictionaries for Colloquial Arabic Dialects

David Graff, Mohamed Maamouri

Linguistic Data Consortium

Primary Goal: Update these 50-year-old dictionaries:

(1) *A Dictionary of Iraqi Arabic: English-Arabic, Arabic-English* (Clarity, et al. 2003 [1965])

(2) *A Dictionary of Moroccan Arabic: Moroccan-English, English-Moroccan* (Harrell and Sobelman 2004) [1966])

(3) *A Dictionary of Syrian Arabic: English-Arabic* (Stowasser and Ani 2004 [1964])

What that goal entails:

- Convert original Latin-based orthography for Arabic words into both a common IPA character set and a useful Arabic script character set.
- Get current native-speaker confirmation for English/Arabic meaning relations and usages.
- Establish "consonantal root" relations among Arabic words where these weren't provided (Moroccan).
- Establish a common basis of reference for both roots and full orthographic forms that supports the recognition of cognates among the dialects, and between each dialect and MSA.
- Where possible, augment the inventory of Arabic look-up words to reflect recent added vocabulary.
- Finish the compilation of the Syrian-to-English dictionary based on extant, unpublished materials (exhaustive archive of hand-written index cards, partial set of printed galley sheets).

Additional goals -- equally important:

- Establish a uniform relational database structure, capable of being extended to cover additional dialects.
- Implement strategies to import existing dictionary contents and supplemental data into the database.
- Provide web-based tools for annotation, query and review of dictionary contents.
- Export completed dictionary content to a standardized transfer format, suitable for use in both hard-copy publication and NLP research: Lexical Markup Framework (LMF) XML.

We gratefully acknowledge the sponsorship of the U.S. Department of Education, whose International Research Study (IRS) Grant No. P017A080044-10 supported our work on this project.

Considerations for applying Arabic-based orthography to Colloquial Arabic dialects:

- Only MSA has an established orthographic standard; no such standard exists for any colloquial dialect.
- Among Arabic speakers, literacy depends on MSA, and builds awareness of the differences and similarities between MSA and a given dialect.
- Differences between MSA and any one dialect are likely to be more constrained / more regular than differences among various dialects.
- Therefore, a pan-dialectical orthographic convention should exploit, as much as possible, the common-core etymology that all dialects share with MSA.

What those considerations entail:

- Arabic letters will have consistent etymological values when used in common-core vocabulary: they represent relations among cognate terms.
- A given letter will have varied phonetic values when viewed across dialects, and may have multiple phonetic values within a given dialect.
- A separate reference must be provided to specify dialect-specific pronunciations for words and phrases (use IPA for this).
- We need to be very careful and thorough about determining etymological relations between MSA and a given dialect, on a word-by-word basis.

MSA	Iraqi	Syrian	Moroc.
ق	q, g (k, ǰ)	ʔ (q, g)	q (g)
ك	k, č (g)	k (č)	k (g)
ث	θ (f, t)	s, t	t
ج	ǰ, č (g)	ǰ, ž	ž (g)

Some Phonetic Correlates of MSA Consonants in Dialects

	Iraqi	Syrian	Moroc.
Total classes	4368	3323	3014
Shared w/MSA	1993	2030	1590
Shared w/others	I/S: 1676	S/M: 1157	I/M: 1433
	I/S/M: 1116		

Tally of Consonantal Root Classes by Dialect
(over 2400 MSA roots are represented in at least 1 dialect)

Relational Database Structure for Bilingual Dictionaries:

- More structure is needed in Arabic-to-English than in English-to-Arabic, to organize lemmas by consonantal roots:
  - Each root relates to one or more lemmas
  - Each lemma relates to:
    - One or more word forms, and
    - One or more English senses
      - Each sense has zero or more example phrases
        - Each phrase comprises reusable word tokens
  - When presenting all the lemmas within a given root, their order should follow established lexicographic conventions, based on part-of-speech and patterns of verb form derivation.
- English-to-Arabic structure is simpler: leave out the "root" and "wordform" layers, use equivalent tables for lemmas, senses, phrases and phrase tokens.
  - But look-up ("headword") entries can include idiomatic phrases, which are subordinate to a prominent lemma used in the phrase.
- In every table containing Arabic (A-to-E word-forms, E-to-A senses, Arabic phrase tokens), store both Arabic script and IPA spellings.
  - Arabic/IPA relations can be context-dependent, irregular, and prone to a variety of annotation errors, posing complex problems for QC.

Annotation tool development: any common web browser plus a stable LAMP framework provide the best environment to implement custom UI's.

Porting DB content to LMF XML:

- Keep the central design strategy:
  - Tags do not bracket arbitrary text content (all tags are "empty").
  - All information is presented as attribute values in the tags themselves.
- The core markup structure for <LexicalEntry> elements is essentially isomorphic with DB table structure.
  - Ordering of elements within a LexicalEntry has arbitrary constraints, but is easy to manipulate via XSLT.
- For ease of visual presentation, the ordering of LexicalEntry elements in the XML stream is significant: lexicographic collation organized into chapters by initial-letter.
- For A-to-E, use "minimal" LexicalEntry elements to present each consonantal root in its proper position, ahead of the lemmas associated with that root.
  - Arabic lemmas that don't involve a Semitic or "productive" root (borrowings, etc) must use the same structure: many "root entries" are actually just "consonant skeletons".